A conversation with Microsoft's Tony Hey

Published 12th December 2006[1]

After thirty years as an academic at the UK's Southampton University, and four years in charge of the UK's e-Science Programme, last year Tony Hey surprised everyone by accepting a post as corporate vice president for technical computing at Microsoft.

Below Hey explains to Richard Poynder why he took the job, and why he believes that his decision to take it is good news for the global research community, and good news for the Open Access Movement.

Unusual academic

RP: *Thank you for making time to speak with me. Let's start with your background? Until joining Microsoft you were an academic who specialised in parallel computing?*

TH: Well, I started out as a particle physicist, and I spent 15 years doing particle physics research, and using the whole gamut of UNIX, and tools like LaTeX.

Then in 1985 I switched to computer science and spent 20 years doing computer science in academia.

RP: *You were based at the University of Southampton. Can you say more about your research there?*

TH: My work was in parallel processing for scientific applications, which when I started was an area that the computer science community had neglected.

RP: *You are talking about high performance computing?*

[1] Updated on 13th December, and 15th December 2006

TH: Right, or supercomputing: those are the names people use. My research, however, was in practical parallel computing. In the mid 1980s, for instance, I worked on the transputer.

RP: *The transputer was a concurrent computing microprocessor developed at INMOS — a company funded by the UK Government right?*

TH: Yes. I worked very closely with INMOS and, with other colleagues, was responsible for developing the transputer, which was in some ways ahead of its time. Today, for instance, we are seeing chips being developed that are very similar to the transputer, but appearing many years later. Had INMOS been properly funded I think the UK would have had a significant impact on the computer industry.

RP: *INMOS was part of the so-called white heat of technology initiative instigated by the British Labour Prime Minister Harold Wilson in the 1960s wasn't it?*

TH: Indeed. The trouble was that when the Conservatives inherited INMOS, they didn't know what to do with it, and sold it off to Thorn EMI. Thorn in its turn didn't understand that it was necessary to invest in the business. So it was a very, very exciting but brief period when the UK seemed to have the courage of its convictions, and just for a minute the country was really competitive in the field.

After the transputer I went into interoperability and portability, and parallel code. One thing I did — with some colleagues — was to write the first draft of the Message Passing Interface [MPI]. This involved a bunch of European and US people meeting every six weeks in Dallas airport hotel, and within a year we had an implementation of MPI that is now an accepted standard around the world.

I was also very keen, by the way, that there should be an Open Source version developed at the same time. So today there are a number of commercial versions of MPI available, plus an Open Source version. There is even a version running on Microsoft products today. I am proud of that.

RP: *In total you spent thirty years at Southampton University?*

TH: Right, although I had 10 years leave of absence.

RP: *And you became head of the School of Electronics and Computer Services department?*

TH: I did, and then I was dean of engineering. So I span the gamut from physics, to computer science, to engineering. In that sense I am an unusual academic.

e-Science

RP: *But your background was clearly ideal for running the UK's e-Science Programme, which you took over in 2001. What is the UK e-Science Programme?*

TH: It was the brainchild of John Taylor when he was running Hewlett Packard's research labs in Europe. He had a vision in which computing would be a utility — a

pay-as-you-go service similar in concept to the pay-as-you-go mobile phone services available today.

RP: *Or the hosting services offered by companies like Google and Amazon (through its S3 service)?*

TH: Exactly. And of course Microsoft now offers such services too — services that are delivered in the cloud: You don't care where they are stored, you just use the services.

Anyway, John was later put in charge of the UK Research Councils, and he found himself working with all the physicists, the chemists, the biologists, and the medics, when they were bidding for money from the government. In fact, it was his responsibility to make those bids.

In doing so, he noticed that a lot of researchers from different institutions were collaborating to do their research, often on an international basis. The particle physics community, for instance, is a genuinely international community, and hundreds of different sites all around the world collaborate with one another.

Other research communities — for example the biologists — might want to collaborate with just a few specific sites: an institute in the UK, say, might want to collaborate with an institute in the US, and an institute in Helsinki. So these three sites would collaborate and share their data.

It was in observing this that John developed his idea of e-Science. Then, when I took over the Programme, it became my task to define it.

RP: *So how do you define e-Science?*

TH: Well, the first point to make is that it's not a science like biology or chemistry. Rather, it is a set of technologies to enable people to collaborate: to share computing, to share data, and to share the use of remote instruments etc. So e-Science is the technologies that allow networked, distributed, collaborative, multi-disciplinary science. It's a very exciting area.

RP: *How does e-Science differ from what in the US is called the cyberinfrastructure. Or are we talking about the same thing?*

TH: Essentially we are talking about the same thing. In fact I had Paul Messina — who was on the US cyberinfrastructure Blue-Ribbon Advisory Panel — on my steering committee; and John Taylor and I were both interviewed by the Blue-Ribbon Panel. So you will see a lot of e-Science ideas in the US cyberinfrastructure report, and you will see a lot of the US cyberinfrastructure report in what we developed. It just happens that in the US they chose another name. Personally, I think e-Science is a much better name than cyberinfrastructure.

RP: *Why?*

TH: Because it emphasises science. The purpose isn't to build roads and infrastructure, but to do science. Of course, e-Science depends on the cyberinfrastructure — the networks, the software, and so on, which we in Europe call the e-infrastructure.

But what is wonderful about the e-Science programme is that it has always been application led.

RP: *You mean that the emphasis has been on what scientists actually want to do, not the technology?*

TH: Exactly. Too often these things are dominated by the technology. And what I really, really liked about the e-Science Programme (and I didn't set it up that way, John Taylor deserves the credit) is that I was only running about 20% of the budget. That is, I ran the core of the Programme, the part that was needed to underpin all the application projects — and the remaining 80% was application-led.

So it was my responsibility to develop the middleware requirements to support the R&D projects, and the applications themselves were directly funded. This meant that the applications were really great, and that is why the e-Science Programme became so visible around the world.

So I believe we had the right idea. The aim was to do serious science, and to tackle next-generation scientific problems.

RP: *Can you give me an example of e-Science in action?*

TH: There are many examples. At one end you have particle physics, where physicists need to share their compute clusters to analyse the data that will soon be generated by the LHC machine in CERN, Geneva. At the other end are astronomers who want to share data from different telescopes all over the world.

One industrial e-Science project, for instance, involves Rolls Royce. Sensor data can be collected from an aero engine while it is in flight: pressure, temperature and vibration data, for instance. This is then sent down for data mining, which involves combining it with the whole maintenance history of that particular engine, and then comparing the results with every single engine of that type that Rolls Royce has ever produced. This means that, if necessary, preventative maintenance can be undertaken. So when the plane lands in New York, for instance, a new part may be waiting to be fitted.

e-Science also offers great benefits to pharmaceutical companies doing drug design. Many of them will have their chemists in one building, their biologists in another building, and their geneticists in another. But in order to collaborate in the workflow for drug design these scientists need to collaborate, and to share data.

All about data

RP: *The logic of e-Science then is that in a networked world it's no longer necessary to work in data silos. You can now share data, and software, and so on?*

TH: Precisely.

RP: *So e-Science is a response to the possibilities of the networked world. But as science becomes more complex I guess researchers need to share more and more data, and in ways they didn't need to before. It's a chicken and egg thing perhaps?*

TH: Sure, and I should maybe stress that I have a slightly different emphasis to some of my friends. For instance, you often hear the word Grid mentioned when people talk about eScience — that is, the word Grid tends to be used synonymously with cyberinfrastructure, or e-infrastructure. Most of the Grid efforts, however, are just about sharing compute cycles. That is what the particle physicists want to do, for instance: just share computers.

But my view is that what most scientists want to do is to share data. And one of the big drivers of e-Science is the fact that over the next five years we will collect more scientific data than has so far been collected in the whole of human history.

RP: *It's more about sharing data than sharing processing power?*

TH: Right. We are going to be deluged with data in almost every field — the particle physicists will have petabytes of information, and almost every other field will have hundreds of terabytes, and petabytes are coming.

The point is that you can't analyse all that data on your workstation: it's just too big, and it is in many different places. Nor are you ever going to move all that data — no matter how fast the bandwidth on the network is. You are just never going to be able to move petabytes around.

For this reason, the so-called supercomputer centres are slightly misnamed. In the future they are really as much data centres, and they will house the computers needed to do the data mining.

RP: *But researchers will need to share supercomputers in order to analyse the data presumably?*

TH: Well, there is a lot of confusion about whether connecting computers creates a supercomputer. In my view it doesn't. With a supercomputer you pay lots of money to have very well tuned, low latency high bandwidth connections between the processes. But when you are running over the Internet you don’t have that, so you do a different type of computing.

One of the wonderful projects we did in the UK — and which I am now funding from Microsoft — was climate*prediction*.net, which is SETI@home-type computing.

RP: *SETI@Home is the distributed computing project run by the US Space Sciences Laboratory, and which uses Internet-connected computers to search for possible evidence of radio transmissions from extraterrestrial intelligence. Anybody can participate by running a free program that downloads and analyses radio telescope data. So instead of having one massively powerful computer, you use the spare processing capabilities of thousands, or millions, of low-powered PCs?*

TH: That's right. And so with climate*prediction*.net you have hundreds of thousands of PCs: your home PC, your school's PC, and so on, to run a climate prediction model. Essentially, they are testing how sensitive the predictions of a climate model of 1950 to 2050 are to the actual input parameters of the model, because obviously we didn't know the input parameters for 1950 with absolute certainty.

Clusters

RP: *So the emphasis in e-Science should be on sharing and managing scientific data, not building supercomputers. In referring to SETI@home-type systems you are, however, also talking about sharing compute cycles?*

TH: Sure, climate*prediction*.net uses lots and lots of cycles from lots and lots of places, and some people call that a supercomputer. In my view, however, it is a lot of cycles, but it is not a supercomputer.

So there is a spectrum from the SETI@home-type model to the supercomputer-type model. Personally, I have never particularly seen much virtue in connecting supercomputers. For most users a cluster of computers is good enough.

RP: *When you say cluster you are talking about having a number of different microprocessors all working together in one place?*

TH: Yes, a cluster is actually a rack of workstations, a rack of PCs in effect. You just buy the boards used in PCs, put them in a rack, and effectively you have a rack of PCs that you can connect together at whatever speed you can afford. A Beowulf cluster, for instance, is a cheap version of a high performance computer.

That, by the way, is one of the nice things about my being in Microsoft: we now have a Windows cluster solution.

RP: *You are referring to the Windows Compute Cluster Server 2003, launched earlier this year?*

TH: Right. The idea is that it will democratise things. So instead of having people in white coats, specialists, telling you how to parallel program we can now put a cluster under your desk and press a button and the code that you were running on your workstation will run 30 times faster on the machine under your desk.

RP: *So the benefit of a cluster is that you get greater computing power to crunch data, but you can do it locally?*

TH: Yes. And so you can do simulation and such like. My point is that most people don't require very high-end computing. For instance, there is a huge amount of buzz about what are called petaflop computers right now, but all that most normal people need are clusters.

RP: *And the Grid, then, is the infrastructure rather than the computers. It's essentially what you call the middleware?*

TH: For me the Grid is just about connecting resources; it's about the middleware to allow you to connect computers and data centres, not so much about compute cycles. And the middleware is part of the cyberinfrastructure.

RP: *So the middleware, or cyberinfrastructure, is needed to enable researchers to access data remotely and, importantly, to share that data and collaborate; to allow groups of researchers in different locations to work together on a particular project?*

TH: Exactly. And so here's another example Richard: We have in the UK a project called Integrative Biology. This is run by a world class group in Oxford who model heart cells — they model the electrical circuits, actually of the chemicals in the individual cells. In Auckland New Zealand, meanwhile, there is another group who do world class mechanical modelling of a beating heart in response to an electrical impulse.

By working together these two groups are able to go from the cell level to the actual beating of the heart. So they can, say, look at gene defects in a heart — where perhaps a chemical that is needed may not be there — and go on to explore heart arrhythmia. So from two isolated research projects you now have the possibility of producing results that are of interest to the pharmaceutical industry.

It's a wonderful project, but the key thing is that it relies on researchers in Oxford being able to share their software programs, share their data, and also maybe connect to the national supercomputer. So e-Science is what allows them to connect up together in order to do those things.

Microsoft

RP: *OK, let's move on to Microsoft. Last year you handed the e-Science Programme over to Hugh Pilcher-Clayton, and joined Microsoft. Why?*

TH: I can understand people's surprise at my doing so. And, let's be quite clear, I was surprised myself. But if you want my absolute honest answer, I simply felt that I had got as far as I could with the e-Science Programme.

RP: *How do you mean?*

TH: The Programme has got a long way down the road, and we made some good progress. But we had got to the stage where we needed the help of the IT companies to make the infrastructure — the middleware — work properly.

RP: *Because the research community can't do this work itself?*

TH: That's right. The cyberinfrastructure is difficult. So I wanted to get Microsoft engaged with the community — and working with the community — and to have it help develop standards.

RP: *But why Microsoft and not, say, IBM?*

TH: Why not indeed. I've used UNIX all my life, and I was much, much closer to IBM. When I started the e-Science program in 2001, therefore, my first action was to call IBM and to ask them to help me; because I didn't think Microsoft was interested in this area.

In the year prior to joining the company, however, I had managed to persuade Microsoft to come to the Grid Forum, and with the support of IBM and others, we produced the first standard of any use to come out of the Forum

RP: *Which standard is that?*

TH: The High Performance Computing Profile, which Microsoft played a key role in getting through.

It occurred to me, therefore, that it would be extremely useful if Microsoft were to continue to collaborate with the Grid community, and to help get interoperable standards agreed in order to build the infrastructure. It was also becoming clear that Microsoft was prepared to do so on a royalty-free basis.

Web Services

RP: *OK. So Microsoft is now interested in playing a part in the high performance and grid computing world. It also wants to help the research community. Indeed, commentators were quick to interpret your recruitment as a signal of Microsoft's intent here. Why has Microsoft become interested in this area?*

TH: One key development has been Web Services, which have become the magic bullet that the IT industry and the computer science people have settled on for doing distributed computing.

RP: *And of course grid computing has started to merge with Web Services, to become what some now call "Grid Services". When we talk about Web Services we are talking about interoperable machine-to-machine interaction over the network?*

TH: That's right. But instead of just trying to create a connection from a resource in one organisation to a resource in another organisation, and not having any control over that external resource, and not knowing whether that resource might change, Web Services allow you to make the resource a specified service.

Moreover, the service is guaranteed by the organisation providing the resource, and the interface is well defined, which means that you do not have to worry about how third parties supply resources. In principle, therefore, it leads to more robust distributed code.

RP: *Another important part of this is that Web Services provide a standard means of interoperating between different software applications running on a variety of different platforms and frameworks. As such, Web Services require non-proprietary standards; standards like the Open Grid Services Architecture [OGSA].*

TH: Yes. So the way I see the middleware infrastructure developing is that it will be based on open standards. This means that there will be a Windows version of a particular piece of software, there will be IBM versions of the software, and there will be Open Source versions. This is the principle that IBM adopted with WebSphere. Do you know about WebSphere?

RP: *Sure. In 1998 IBM announced that it would start selling and supporting the Open Source web server Apache, which it built into its WebSphere product. This was widely viewed as a major breakthrough for the Open Source Movement.*

TH: The point is that today there is an Open Source implementation of Web Services, and there is IBM's commercial product called WebSphere and both of them interoperate.

RP: *So in order to be able to play in the distributed computing world, therefore, Microsoft decided to embrace the less proprietary software environment that Web Services require; an environment in which many different platforms will interoperate?*

TH: Well, in distributed systems it is essential that the various Web Services are open. And the distributed system we are talking about here is essentially the cyberinfrastructure.

So my vision is that when we have this software infrastructure in place there will be Open Source versions of it as well as propriety versions. Linux will always be here — I absolutely agree — but we will have heterogeneous systems, with bits of Windows, bits of Linux, bits of Open Source, whatever, all interoperating. That is the way it is going to be.

What this also means is that there will be genuine competition, which is good for the whole community.

RP: *I'm conscious that there was a lot of suspicion when work began on the Web Services infrastructure, not least because IBM and Microsoft began patenting the interfaces.*

TH: That was because they knew that if they didn't then somebody else would, and since IBM and Microsoft have a lot of money they would have been targeted by whoever got those patents.

The important point, however, is that both IBM and Microsoft have jointly said that these interfaces are available for anyone to use, royalty free. And as a sign of its commitment to this principle, in September Microsoft issued what it calls the Open Specification Promise.

RP: *Essentially this is a promise not to sue?*

TH: It says that you can do anything you like with any of the technologies covered by the promise. You don't have to mention Microsoft, you don't have to sign anything, and you don't even need to communicate with Microsoft. You just have to agree to the terms in order to benefit from the promise.

In return we promise we won't sue you. It is wonderful promise. I couldn't have written it better myself! And IBM has done something similar thing.

The truth is that Web Services have really been driven, and become successful, because IBM and Microsoft agreed to agree. Bill Gates stood up with Steve Mills and essentially they said, "Our Web Services will interact with each other, and with Open Source Web Services."

RP: *Specifically, we are talking about Web Services specifications like SOAP, WSDL and WS-I?*

TH: We are. What Microsoft has also done is to commit not to sue anyone over the XML Reference Schema in Office 2007 — which is just about to be released.

Open XML

RP: *You are referring to* Open XML*?*

TH: Right. The formats used in Office 2007 applications like PowerPoint, Excel Spreadsheets and Word are defined in Open XML.

RP: *Can you say more about the commitment that Microsoft has made with regard to Open XML?*

TH: It is a commitment that says that you can use the Open XML schema in connection with other applications without fear of being sued.

So, for instance, you might want to use the Open Source product Open Office to manipulate documents produced by a Microsoft application. In doing so, however, you would probably be violating some of our patents on some of the technologies required to process the Open XML format. So we covenant not to sue you for doing so.

In addition, Microsoft has submitted Open XML to the standards body ECMA. So it is not even under our control any more.

RP: *Yes, and I understand that it has just been certified by ECMA. However many predict that over time it will be the Open Document Format* [ODF] *that becomes the standard, not Open XML. IBM even appears to believe that Open XML is redundant.*

TH: This is a very interesting point Richard. The fact is that the format for Word is a *de facto* standard. So while the Open Document Format is very elegant and simple, there are billions of documents in Word format out there.

What we have done, therefore, is to publicly specify what the Open XML format is. In addition, we are sponsoring an Open Source translation effort — to allow people to translate between Open XML and ODF.

RP: *How will that work*?

TH: It means that if you want to use ODF you can press a button and it will produce an ODF document. It just won't be so rich, and it won't be as beautiful as a Word document.

RP: *Microsoft is not providing native support for ODF however. Might that not have been a more useful thing to do?*

TH: As I say, Microsoft has supported an Open Source activity to produce a translator. One again, the point to bear in mind is that in the future there are going to be many different software standards. The important thing, therefore, is to ensure that those standards are mutable, and translatable.

As a further sign of Microsoft's commitment to openness it has also developed several Open Source licences of its own, which we call Shared Source Licences.

RP: *Are you saying that some of Microsoft's software is now Open Source?*

TH: I am saying that some of our code is available for people to view, and in some cases modify. And if you teach Windows on an operating system course we will deliver the code to you, and you can use it to compare the source code of Windows with, say, the source of UNIX.

RP: *So Microsoft has changed: It is more open?*

TH: I believe so. We now have open Web Services standards, we have open standards for Office, and you can get the source code of Windows.

RP: *So how would you characterise Microsoft's attitude to Open Source today: does it love Open Source, does it like Open Source, or is it more that it has learned to live with Open Source?*

TH: Microsoft can live with Open Source. And for all the reasons we have discussed you will perhaps see why it was not hard for me to join the company. What has proved an unexpected bonus for me, however, is that I am now also working on Open Access.

Open Access

RP: *You support Open Access?*

TH: I'm passionate about Open Access.

RP: *That makes sense. The premise of e-Science, presumably, is that scientific information needs to be freely available?*

TH: Right. As I said, the key part of the cyberinfrastructure — although it is not always mentioned — is not so much the network and the middleware, but enabling Open Access, both to the scientific literature and to scientific data. So yes, the assumption is that there will need to be some form of Open Access. This, however, is now inevitable, and you can see it beginning to happen. You are aware of the Cornyn-Lieberman Bill in the States?

RP: *The Federal Research Public Access Act of 2006? Yes.*

TH: There is also an EU proposal, and a number of initiatives from the UK Research Councils. And of course The Wellcome Trust has already introduced a self-archiving mandate. So it is only a matter of time.

RP: *It is interesting that although Open Access appears inevitable in a digital networked world, it has its roots in the pre-Web era. The so-called serials crisis, for instance, dates from at least the 1980s, and probably somewhat earlier.*

TH: Indeed. There has been a worldwide crisis in the price of journals for some while — a problem that first came to my attention when I was at Southampton. I am an editor of a Wiley journal on parallel computing, and I discovered that the University doesn't take the journal.

I had my own copies of the journal, and so was able to share them with my group. But it was only when I looked into why the University didn't have a subscription that I understood the reason why. Every year I was getting a note from the library asking which journals could be cancelled: They simply couldn't afford all the journals faculty need any longer.

RP: *The hope is that Open Access will remove these financial barriers, and allow anyone who needs to have access to the scholarly literature to have it. Right now the favoured approach is to mandate researchers to self-archive their published papers in their institutional repository.*

TH: Exactly. So the vision I have — and it is shared by colleagues at Microsoft like Jim Gray, and many people outside Microsoft too — is that there will be a kind of digital library in the sky. This will be composed of subject repositories like arXiv and the hundreds of institutional repositories being created.

The challenge will be to connect all these repositories together, and to make them interoperable and searchable. That is the big vision we have.

RP: *Microsoft is committed to facilitating Open Access as part of its mission of helping to build the infrastructure for e-Science then?*

TH: We are certainly helping. And, as I said, while I didn't expect it when I joined the company, I was amazed to discover that Microsoft was already supporting Open Access.

RP: *In what way?*

TH: It turns out, for instance, that Microsoft has been supporting the deployment of portable PubMed Central. Do you know about portable PubMed Central [pPMC]?

RP: *PubMed Central I know — it's the US National Library of Medicine's free online repository of biomedical and life sciences literature. What is portable PubMed Central?*

TH: It's an XML-based project that enables the creation of portable versions of PubMed Central. The Wellcome Trust and the British Library are currently putting up a version in the UK as we speak.

RP: *OK, you are talking about creating national mirrors of PubMed Central, the first of which is UKPMC?*

TH: Right. Microsoft has been working with the US National Center for Biotechnology Information [NCBI] in their development of "portable PubMed Central" — now PMC International — since early 2004. One of the joint activities involved making use of the editing tools in Microsoft Word to author NML tagset XML. Another involved using XML with an SQL Server database.

We are also helping to deploy PMC International in South Africa, in Japan, in China, and in Italy. So people all around the world will be use this valuable Open Access library of medical literature.

EPrints

RP: *And one of your first actions on arriving at Microsoft was to provide funding for porting the institutional repository software EPrints to Windows. Currently EPrints is only available on the GNU/Linux platform?*

TH: Yes. EPrints is fine, but it is difficult to configure, and librarians are telling us that they want it on Windows. So we have agreed to help develop a Windows version.

RP: *This is because many librarians want to create institutional repositories on their existing Windows platform?*

TH: And they want it to be easier.

RP: *Could it be easier? Open Access advocates like Steven Harnad maintain that using EPrints it is possible to put up an institutional repository in 20 minutes without difficulty.*

TH: Stevan belongs to a computer science department, as do the EPrints developers at Southampton. They don't like it when university librarians tell them we need this feature, or we need that feature. Consequently, librarians have a very tough time using EPrints.

My wife is a librarian, and she has created an institutional repository using EPrints, so I take a personal interest in this. Are you aware of Tardis?

RP: *Tardis was a pilot institutional repository at Southampton University.*

TH: That's right. The project was led by the Southampton oceanography centre, but it was my wife and her colleagues who built it, using EPrints. They did a great job, but it was extremely difficult. Moreover, at the end of the day the result was a pragmatic one.

RP: *Pragmatic?*

TH: Well, Open Access advocates would like repositories to be all full-text. But in Tardis many of the records are only abstracts that refer you to the publisher's web site, where you have to pay for the full-text.

RP: *Loading full-text also raises licensing issues I guess?*

TH: That's right. That is the reality of the situation we are in at the moment. Unfortunately, Open Access advocates don't always understand the complexities, so I take a slightly different view to Open Access than they do.

RP: *Nevertheless, you believe Open Access is inevitable?*

TH: I do. And I am assuming that eventually all the content in repositories like Tardis will either be full-text, probably with delayed Open Access like PubMed Central [most papers in PubMed Central are archived after a six month embargo] or, as OA advocates would like, immediate Open Access on publication.

RP: *Do you yourself believe that immediate Open Access is the optimum approach?*

TH: On such fine details as embargoes I tend to be neutral but, as I say, some form of Open Access is inevitable.

RP: *Has the funding for developing a Windows version of EPrints begun?*

TH: I have certainly signed the cheque; but I don't know whether they have done any work yet.

GPL vs. BSD

RP: *As I understand it, the Linux version of EPrints is licensed under the Free Software Foundation's General Public Licence [GPL]. You are insisting, however, that — as a condition of funding — the Windows' version must be licensed under a BSD (Berkeley Software Distribution) licence?*

TH: Well, clearly Southampton owns the copyright to the software, so they can do what they want. I don't want to dictate what licence they choose, but the one thing I cannot fund is a GPL-type licence. I cannot support a licence that denies the existence of a software industry.

RP: *Both the GPL and BSD are Open Source licences. What you object to, I guess, is the so-called "viral" character of the GPL, which requires that any software derived from GPL-licensed code must be distributed under the same copyleft terms as the original software — thereby ruling out the possibility of using the code to create a proprietary product. By contrast, anyone can use BSD-licensed code to develop a proprietary version. But is it fair to say that the GPL denies the existence of the software industry?*

TH: Well, with the GPL you can only provide support services in the way that Red Hat does. Does that deny the existence of a software industry? I'm not sure, but it would certainly make it more difficult for a popular software product like Matlab to be developed.

RP: *Do you object to the GPL because Microsoft has a policy of not supporting it, or is it that you personally do not like it?*

TH: It's me personally. Actually, I don't know what the Microsoft policy is. I know Microsoft's Steve Ballmer has been talking to Red Hat and so on, but I am not going to be able to comment in depth about Microsoft's policy.

RP: *So you personally view the GPL as a bad thing?*

TH: I will always respect somebody who wants to go the GPL route, that's fine, they can do that, it is their choice. But it is my belief that if you want to encourage innovation in science and engineering, and you want to do so by developing Open Source software, then a BSD-style licence is the most appropriate way to do so.

RP: *Do you have personal experience of the BSD?*

TH: I do. The Open Source version of the Message Passing Interface was developed under a BSD licence. And when I was dean and head of department at Southampton I helped Wendy Hall — who is now the chair of the department — spin off her own company using the Microcosm code, which had been developed with funding from JISC.

It is because the Microcosm software was developed using a BSD licence that people at Southampton were able to put some effort into making a proprietary version and then form a company around it.

RP: *So it is important to you that EPrints is licensed under a BSD-style licence because that allows anyone to use the code to develop a commercial product?*

TH: For example someone at Southampton, yes.

Not dogmatic

RP: *But I wonder if anyone would want to? Stevan Harnad argues that EPrints is just a short-term tool, a kind of lever to enable Open Access, not a long-term product from which money can be made.*

TH: That might be true. But listen Richard, I am not dogmatic about this As I said, I am just responding to what the librarians say, which is that they would be more interested in EPrints if they had a Windows version; and I can see that there would be more take-up if there was a Windows version of EPrints. That's all; it's as simple as that.

At the same time I also hope that some of my colleagues in universities will think about how they can generate businesses for the next generation. And that seems to me to be what universities should be doing.

RP: *You are saying that universities have a role in fostering the software industry?*

TH: I happen to believe that the role of universities is to generate jobs, and industry, and innovation for their country.

RP: *Some might argue that it isn't the job of academics to start businesses. But perhaps that is a different conversation.*

TH: Yes. It is. And I think it is different in the States. There the National Science Foundation (NSF) has always taken the view that universities should devote themselves to upstream, basic research, and that they should operate very differently from industry.

In the UK, however, both the Government and the Research Councils take the view that they put all this money into research, so it is not unreasonable that some of it should go to helping UK industry.

That is a view that I endorse. And that is why I believe it is important for universities to spin off companies, to help innovation, and to start jobs for people in the UK, and indeed Europe.

I am in favour of Open Source in this situation because universities are funded from public money and so the code they write should be open. But I prefer a BSD-style Open Source licence because it allows academics to start companies.

Going to change

RP: *You say that you are passionate about Open Access. I'd like to raise with you a point made by Paul Ginsparg, the founder of arXiv.*

TH: OK.

RP: *Speaking in an interview he gave earlier this year with Educause's Matt Pasiewicz, Ginsparg expressed some disappointment that, despite the potential of the Internet for revolutionising scholarly communication, researchers cannot yet create dynamic documents capable of exploiting the capabilities of the digital network. This, he said, is because the tools available to them still only allow them to produce static documents more suited to the print world. And the reason for this, had added, is that large corporations like Microsoft are more interested in satisfying their shareholders than in meeting the needs of their customers. Indeed, with over 90% market share in the Office application space, there is little or no incentive for Microsoft to innovate in ways that would benefit researchers. Consequently, he said, "Right now we are still stuck in the same place we were a decade ago." Does he have a point?*

TH: Well, with Office 2007 and Open XML you can do everything that Paul wants. I know Paul: he and I come from the same community of theoretical physics. And a year ago I went to Cornell University to speak to him. Following that meeting we now have a joint project to look at arXiv, and to explore ways in which Open XML can provide what Paul wants.

RP: *Yes, he mentioned that he was talking to Microsoft. Nevertheless I'm not sure he agrees that Open XML will solve the problems he has identified. Or that Microsoft will help. It was clear, he said, that there are "a number of institutional issues" at Microsoft's end.*

TH: Believe me, with Open XML it is now possible to do many things that you couldn't do in the past.

RP: *I guess you are saying, "That was the past. It is all going to be different in the future."*

TH: Come on Richard. The whole IT industry is constantly changing. There are all sorts of interesting things coming out from all sorts of companies. Web 2.0 offers a wonderful set of tools: wikis, blogs, RSS feeds, and so on.

So I agree with Paul: there is a really wonderful chance to change scholarly publishing for the benefit of research. Web 2.0 makes everything different.

RP: *So Microsoft is keen to help researchers improve scholarly communication?*

TH: Absolutely. And I passionately believe that the whole nature of research publishing is going to change. What, for instance, is a research paper? After all, today researchers don’t need a volume of a journal; all they want is a particular paper.

And then there is refereeing: I believe peer review is very important. But Paul Ginsparg puts up documents that are not peer reviewed on arXiv, and that seems to work for physicists.

RP: *Peer review is set to change?*

TH: Maybe. Most people would like some form of reviewing to take place, but how that is done may differ. BioMed Central, for instance, has a service that it calls Faculty of 1000, where 1,000 academics comment on what they consider to be the most interesting biomedical papers in the last month. That is another form of refereeing. So Tony Hey rated this paper 4 out of 5, and he also liked these ones too. It's similar to the way Amazon works.

Ease of use

RP: *How then do we characterise what Microsoft can bring to e-Science, and to Open Access?*

TH: Microsoft can make these things more easy to use, and provided it does that in an interoperable way, and in a way that doesn't get up the nose of the community, that is good for everyone.

RP: *What do you mean when you say that Microsoft can make these things more easy to use?*

TH: I mean that Microsoft can provide the academic community with what it has already provided for the consumer and business markets: ease of use.

The way it currently works in the academic community is that graduate students are considered cheap labour. So, for instance, in chemistry research groups one graduate student is always sacrificed to be the UNIX expert. But if you think about it, that person wanted to do research in chemistry, not become a UNIX expert.

So we can help by making it possible for that graduate student to do chemistry, not UNIX programming.

RP: *Because if Microsoft can make e-Science as easy to do as, say, writing a document in Word, then these graduate students can be freed up to get on with their research?*

TH: Precisely. Today, when you turn on your PC you don't have to worry about configuring the TCP/IP stack in order to connect to the Internet. You used to have to do that, but now it is routine.

What we need to do, therefore, is to make it just as easy for a bunch of chemists, or biologists, or physicists located in various different places, to share and collaborate: to enable researchers in Manchester, for instance, to collaborate with researchers in Cambridge and/or Southampton, or Bristol.

They might want to share data, for instance, and they might want to do so in a secure environment — because they are working on scientific papers and they don't want anybody else to see the data. Our task, therefore, is to help them to create their own little virtual organisation, and to be able to do so quickly and easily.

As I said, we made a good start on these things with the e-Science Programme, but we have not yet succeeded. What we need now is for the IT industry to play a part in the process.

RP: *And what role do you see Microsoft playing in the Open Access space?*

TH: We've discussed some of the things we are doing. In addition, I have been working with the Mellon Foundation, with Herbert Van de Sompel, and with Carl Lagoze, looking at how we can create interoperable repositories that are searchable at a more fundamental level than is currently possible with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and using Dublin Core.

RP: *Some Open Access advocates might argue that the current system is good enough.*

TH: Well, as I said, not everyone understands the complexities. I personally have no doubt that more work is needed, and I would like to help with that. For that reason I provided funding for a workshop held by the Mellon Foundation to demonstrate interoperability across different repository software.

What is currently stupid is the way in which it is assumed that DSpace, or EPrints, or Greenstone, or Fedora, will become the one solution. It's not going to be winner takes all; it is going to be a heterogeneous world. So there is more work to be done on interoperability.

I am also hoping to work with the publishers, and with the Open Access community, to see if we can find out what is wanted on campuses.

Like everyone else, publishers need to recognise that the Web has changed everything for scholarly communication, so we have to find a model where they still have a business, but the model is also welcomed by the academic community, and not considered to be an impossible burden.

For that reason I am talking to people like *Nature*, and I am trying to do a project with Wiley, looking at Open Access, and looking at new business models that are not so unfair on universities.

Pretty altruistic

RP: *What does Microsoft hope to get out of supporting e-Science and Open Access?*

TH: We need to learn. Until now Microsoft has not been working with the research community. So it needs to learn how to do that, and it needs to learn that academics are intelligent, that they have their views, and that they want choice: they can't be forced.

It's not like a company, where you can say that everyone is now going to use Linux, or everyone is going to use Windows. That is not the way the academic research community works: they will make choices about what they are going to use, and we have to respect that.

So we want to talk to researchers, to listen and learn, and so understand how we can offer things to the community that it feels are acceptable, and to provide things they want. We won't get any market share unless we do.

RP: *So at the moment you are primarily in listening mode. Sooner or later, presumably, Microsoft will want to sell some products to researchers?*

TH: You may not realise it but academics today can get all of Microsoft's software — to use for non-commercial research purposes — for a few hundred dollars. They can get anything they like from Microsoft that way, so the software is very nearly free for universities. To talk of selling into the academic community, therefore, is a strange term to use.

No, the sales will come from the pharmaceutical industry, from the oil industry and from the manufacturing industries, who all want supported products; I don't think we are looking to the academic community as a major source of revenue.

RP: *So Microsoft's support for researchers is purely altruistic?*

TH: Clearly it isn't entirely altruistic: we are a company. What I hope is that Microsoft can learn about the community, and become a trusted partner. And I mean partner, not a dominant player.

RP: *I'm struggling to understand what is in it for Microsoft if the company is not looking to sell products to researchers?*

TH: I've told you: We are interested in producing things that enable industry to do its research; to provide tools and technologies that will deliver benefits to people in, say, the pharmaceutical industry.

RP: *If you want to develop products for pharmaceutical companies wouldn't it make more sense to talk to them directly, not to the academic community?*

TH: We are talking to them too. But my task is to deal with academia, which I see as a place where we can experiment, and where we can learn, and work out what works and what doesn't work. In doing so I believe we can benefit academic research.

We can develop Grid standards, for instance, and make the whole middleware infrastructure much more stable. And we can help make this whole "repository in the sky" system work.

Of course this will have commercial implications, but for the academic community there are very few strings attached to working with us, so far as I can see. So I guess it is a pretty altruistic approach.

A statement you can trust?

RP: *Microsoft now accepts that we live in a heterogeneous distributed world where many different types of software will interoperate and co-exist by means of open standards?*

TH: That's right. As I said, I expect that the university environment will have Windows, Open Source, and UNIX of various varieties, all capable of interoperating. That is how it should be.

RP: *There is nevertheless a lot of scepticism as to whether Microsoft is able to abandon its proprietary mindset?*

TH: And I think you are one of the sceptics!

RP: *That's true. I am also conscious that in a c|net article published last year that quoted you commenting on Microsoft's new commitment to openness, an analyst from illuninata called Jonathan Eunice also voiced scepticism. Responding to your comments, he said, "Tony's endorsement of open standards is quite interesting and I think significant, but I don't think it's a statement you can trust. It's open as far as open benefits Microsoft."*

TH: [laughs] I saw that too…

RP: *To support his scepticism Eunice pointed out that Microsoft has chosen not to implement Globus Toolkit — the Open Source toolkit for building computing grids — on Microsoft's new cluster machine.*

TH: That's right. And he went on to say that the reason I was stressing the importance of data was because we weren't serious about this.

But as I have explained to you, from my e-Science perspective I believe data to be the key issue — data federation, data integration, and data mining. These things are central. The core challenge is how you combine things like genomics data — which is gene sequences — with microarray data; how you combine a two-dimensional image with three-dimensional protein data; and how you figure out, for instance, what a drug is.

RP: *Which brings us back to where we started: the importance of data?*

TH: It does. As I say, my belief is that this is all about data. And this is based on a genuine belief, not because I want to bash Globus. Globus are actually good friends of mine, and when I was in the UK I put together a Globus Alliance with the Edinburgh folks that work with Globus. So I have nothing against Globus.

RP: *So why not implement it on Microsoft's Cluster Server?*

TH: Because there were some false starts with the Grid Community. If you go to the Global Grid Forum (GGF), for instance, you wonder why it has produced nothing in

five years. Well, that is because they had two false starts with Web Services: Instead of just building on what was there they tried to define new services in the Grid community, and they expected the Web Services community to adopt them. The trouble with that approach is that the Grid Community is like a pimple on the back of an elephant compared to the Web Services community. It just doesn't have the power to force new Web Service standards.

When it became clear that the first effort had failed, GGF and Globus went in another direction, which was to use IBM's WSRF.

RP: *So they got there in the end?*

TH: It isn't very different from what I think will eventually emerge, but companies like Microsoft, Sun and Intel did not support WSRF. The point is that you really have to have standards that are supported by the vast majority of the IT industry, and the community, and you don't want to build standards for grids on things that are still being defined.

RP: *OK, my final question then: in both the Open Source and Open Access environments the business model is predominantly one of selling services, not products. That seems to be the direction IBM is moving in the Open Source space, and it is the model adopted by Open Access publishers like Public Library of Science and BioMed Central — who sell publishing services to authors, rather than journals to readers. Is that the future direction for Microsoft too: will it migrate its business model from one based on selling products, to one in which it primarily sells services?*

TH: As you know, we are just releasing the Windows Vista and Office 2007 products. But we also have the Windows Live and Office Live services, and we have Xbox Live. So yes, we are offering Live services. And in the future it is likely that you will use more and more data that is hosted somewhere else in the cloud.

But while I am sure we will see lots of experimentation out there, it is difficult to predict exactly what the world will be like in the future. My expectation is that there will be an interesting mix of services offered, and you will sometimes find it convenient just to buy a service, but there will always be a place for software products.

RP: *OK, thanks for your time.*

If you would like to republish the article on a commercial basis, or have any comments on it, please email me at richard.poynder@btinternet.com.

Please note that while I make this interview freely available to all, I am a freelance journalist by profession, and so make my living from writing. To assist me to continue making my work available in this way I invite anyone who reads this interview to make a voluntary contribution. I have in mind a figure of $8, but whatever anyone felt inspired to contribute would be fine. This can be done quite simply by sending a payment to my PayPal account quoting the email address *richard.poynder@btinternet.com*. It is not necessary to have a PayPal account to make a payment.